Efficient Large-Scale Video Understanding in The Wild
Yi Zhu and Shawn Newsam  
yzhu25, snewsam@ucmerced.edu

**Motivation:**
- Enormous explosion of user-generated videos, containing a wealth of information. However, it would take forever to manually annotate the data and make use of them.
- Need for video-based applications, like video search, video highlighting, video surveillance, etc. Recent trending topics in computer vision include video action/event recognition.

**Objective:**
Efficient large-scale video understanding in the wild. Specifically,
- Better encoding of static frame appearance information (image classification)
- Better description of short term motion between adjacent frames (action recognition)
- Better exploration of temporal structure and extraction of video/clip-level features instead of frame-level features (event recognition)
- Better fusion of different channels of information (multi-modal learning)
- Better identification of spatial patterns geographically according to video metadata information like geo-coordinates. (smart city)

**Past work:**
1. Depth2Action: Exploring Embedded Depth for Large-Scale Action Recognition (ECCV 2016)
   - This paper performs the first investigation into depth for large-scale human action recognition in video where the depth cues are estimated from the videos themselves

2. Efficient Action Detection in Untrimmed Videos via Multi-Task Learning (WACV 2017)
   - This paper studies a multi-task learning framework that performs the three highly related steps of action proposal, action recognition, and action localization refinement in parallel instead of the standard sequential pipeline that performs the steps in order

**Work in progress:**
Hidden Two-Stream Networks for Action Recognition
- We present a novel CNN architecture that implicitly captures motion information for action recognition. Our method is 10x faster that a conventional two-stage approach, does not need to cache flow estimates, and is end-to-end trainable.

**Code and models available:**
https://github.com/bryanyzhu/Hidden-Two-Stream

---

**Diagram:**

   - This paper studies a multi-task learning framework that performs the three highly related steps of action proposal, action recognition, and action localization refinement in parallel instead of the standard sequential pipeline that performs the steps in order

   - Our parallel model is more robust than its sequential counterpart when limited training data is available.

---

**Tables:**
- | Model | Train | Val | Test |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth2Action</td>
<td>18.0</td>
<td>12.0</td>
<td>6.0</td>
</tr>
<tr>
<td>Our Parallel</td>
<td>12.0</td>
<td>8.0</td>
<td>6.0</td>
</tr>
</tbody>
</table>

---

**Code and models available:**
https://github.com/bryanyzhu/Hidden-Two-Stream